The 15th Conference of the Association for Machine Translation in the Americas

2022.amtaweb.org

PROCEEDINGS

Workshop on Corpus Generation and Corpus Augmentation for Machine Translation

Organizer: John E. Ortega

Proceedings of the First Workshop on Corpus Generation and Corpus Augmentation for Machine Translation (CoCo4MT) at the AMTA 2022 Conference

Organizers
John E. Ortega[1,2] john.ortega@usc.es
Marine Carpuat[3] marine@umd.edu
William Chen[4] williamchen@cmu.edu
Katharina Kann[5] katharina.kann@colorado.edu
Constantine Lignos[6] lignos@brandeis.edu
Maja Popopvić[7] maja.popovic@adaptcentre.ie
Shabnam Tafreshi[3] stafresh@umd.edu

[1]New York University
[2]University of Santiago de Compostela - CITIUS
[3]University of Maryland
[4]Carnegie Mellon University
[5]University of Colorado Boulder
[6]Brandeis University
[7]ADAPT Centre

1 Aim of the workshop

The first workshop on corpus generation and corpus augmentation for machine translation (CoCo4MT) sets out to be an original workshop centered around research that focuses on corpora creation, cleansing, and augmentation techniques specifically for machine translation.

We hope that submissions will provide high-quality corpora that is available publicly for download and can be used to increase machine translation performance thus encouraging new dataset creation for multiple languages that will, in turn, provide a general workshop to consult for corpora needs in the future.

2 Workshop scope and details

It is a well-known fact that machine translation systems, especially those that use deep learning, require massive amounts of data. Several resources for languages are not available in their human-created format. Some of the types of resources available are monolingual, multilingual, translation memories, and lexicons. Those types of resources are generally created for formal purposes such as parliamentary collections when parallel and more informal situations when monolingual. The quality and abundance of resources including corpora used for formal reasons is generally higher than those used for informal purposes. Additionally, corpora for low-resource languages, languages with less digital resources available, tends to be less abundant

and of lower quality.

CoCo4MT sets out to be the first workshop centered around research that focuses on corpora creation, cleansing, and augmentation techniques specifically for machine translation. We accept work that covers any spoken language (including high-resource languages) but we are specifically interested in those submissions that are on languages with limited existing resources (low-resource languages) where resources are not highly available. Since techniques from high-resource languages are generally statistical in nature and could be used as generic solutions for any language, we welcome submissions on high-resource languages also.

The goal of this workshop is to begin to close the gap between corpora available for low-resource translation systems and promote high-quality data for online systems that can be used by native speakers of low-resource languages is of particular interest. Therefore, It will be beneficial if the techniques presented in research papers include their impact on the quality of MT output and how they can be used in the real world.

CoCo4MT aims to encourage research on new and undiscovered techniques. We hope that submissions will provide high-quality corpora that is available publicly for download and can be used to increase machine translation performance thus encouraging new dataset creation for multiple languages that will, in turn, provide a general workshop to consult for corpora needs in the future. The workshop's success will be measured by the following key performance indicators:

- Promotes the ongoing increase in quality of machine translation systems when measured by standard measurements,
- Provides a meeting place for collaboration from several research areas to increase the availability of commonly used corpora and new corpora,
- Drives innovation to address the need for higher quality and abundance of low-resource language data.

Please feel free to review the official workshop website: `https://sites.google.com/view/coco4mt` for more information and details.

3 Invited Speakers (listed alphabetically by first name)

We are happy our dear colleagues Ankur Parikh, Jörg Tiedemann, Julia Kreutzer, Graham Neubig, and Maria Nadejde have prepared talks on five important topics for CoCo4MT 2022.

3.1 Ankur Parikh, Google Research

Ankur Parikh is a staff research scientist at Google NYC. His research interests are in natural language processing and machine learning with a recent focus on high precision text generation. Ankur received his PhD from Carnegie Mellon in 2015 and has received a best paper runner up award at EMNLP 2014 and a best paper in translational bioinformatics at ISMB 2011.

3.2 Graham Neubig, Carnegie Mellon University

Graham Neubig is an associate professor at the Language Technologies Institute of Carnegie Mellon University. His research focuses on multilingual natural language processing, natural language interfaces to computers, and machine learning methods for NLP, with the final goal of every person in the world being able to communicate with each-other, and with computers in their own language. He also contributes to making NLP research more accessible through open publishing of research papers, advanced NLP course materials and video lectures, and open-source software, all of which are available on his web site.

3.3 Jörg Tiedemann, University of Helsinki

Jörg Tiedemann is professor of language technology at the Department of Digital Humanities at the University of Helsinki. He received his PhD in computational linguistics for work on bitext alignment and machine translation from Uppsala University before moving to the University of Groningen for 5 years of post-doctoral research on question answering and information extraction. His main research interests are connected with massively multilingual data sets and data-driven natural language processing and he currently runs an ERC-funded project on representation learning and natural language understanding.

3.4 Julia Kreutzer, Google Research

Julia is a research scientist at Google Research, Montreal, where she works on improving machine translation. She is generally interested in the intersection of natural language processing (NLP) and machine learning. In her PhD (Heidelberg University, Germany) she investigated how reinforcement learning algorithms can be used to turn weak supervision signals from users into meaningful updates for a machine translation system.

3.5 Maria Nadejde, Amazon

Maria is a Senior Applied Scientist at Amazon AWS AI working on improving quality and customization of Amazon Translate. Before joining Amazon, Maria was an Applied Research Scientist at Grammarly developing deep learning applications that enhance written communication. She obtained a PhD in Informatics from the University of Edinburgh on the topic of syntax-augmented machine translation.

3.6 Other speakers and guests

CoCo4MT decided to create a panel that includes several other researchers and notable speakers in order to provide collaboration amongst those wanting to assist with low-resource language approaches for corpora augmentation and generation. These speakers are to be announced in future (post-edited) version of the proceedings.

4 Program Committee (listed alphabetically by first name)

- Amirhossein Tebbifakhr, University of Trento
- Anna Currey, Amazon
- Ayush Singh, Northeastern University
- Barry Haddow, University of Edinburgh
- Bharathi Raja Chakravarthi, National University of Ireland Galway
- Beatrice Savoldi, University of Trento
- Constantine Lignos, Brandeis University
- Eleftheria Briakou, University of Maryland
- David Adelani, Saarland University
- Jasper Kyle Catapang, University of Birmingham
- John E. Ortega, University of Santiago de Compostela - CITIUS
- Jonathan Washington, Swarthmore College

- Jonne Sälevä, Brandeis University
- José Ramom Pichel Campo, University of Santiago de Compostela - CITIUS
- Katharina Kann, University of Colorado Boulder
- Kochiro Watanabe, The University of Tokyo
- Liangyou Li, Huawei
- Maja Popopvić, ADAPT Centre
- Maria Art Antonette Clariño, University of the Philippines Los Baños
- Marine Carpuat, University of Maryland
- Pablo Gamallo, University of Santiago de Compostela - CITIUS
- Patrick Simianer, Lilt
- Rico Sennrich, University of Zurich
- Rodolfo Joel Zevallos Salazar, Universitat Pompeu Fabra
- Shabnam Tafreshi, University of Maryland
- Shantipriya Parida, Idiap Research Institute
- Surafel Melaku Lakew, Amazon
- William Chen, Carnegie Mellon University
- Xing Niu, Amazon

Contents